MARYVILLE UNIVERSITY

Predictive Modeling

Week 5

ACSC412/512

Fall 2017

Richard Xu

What were covered last time

1. Classification model, example
2. Summary of GLM
3. Modeling in the real world (specification, variable selection, validation)

What we are going to discuss today

1. Homework 3/4
2. Tree model Introduction
3. Regression & Classification with examples
4. Clsutering– K-means and hierarchical; insurance application
5. Project

Homework3

Problem1

1. pM1 <- glm(AvgClaim ~ factor(S) + T, data=x, family=Gamma(link="log"), weights=exposure)
2. Major issue: target, weights, offset
3. Exposure? Link function?
4. Necessary steps

Problem2

1. miM4 <- glm(numclaims ~ offset(log(exposure)) +veh_value+veh_body+veh_age+areaGP+factor(agecat)+ veh_value:veh_age, family = poisson(), data = x)
2. Target, offsets, weight
3. Exposure/payment?
4. Necessary steps

Homework4

Problem1

1. glm(LapseR ~ …, family=binomial, weights=exposureN,…)

   glm(cbind(LapseN, ExposureN-LapseN)~…,famile=binomial,…)

   glm(LapseN ~ offset(log(ExposureN))+…,famile=possion,…)

2. Same problem for target, offset, and exposure

3. Same target variable for comparison (RiskClass?)

Problem2

1. Simple logistic mode

glm(DamageProb ~ …, family=binomial,…)

glm(cbind(DamageIndex, 20-DamageIndex)~…,famile=binomial,…)

Offset and weight

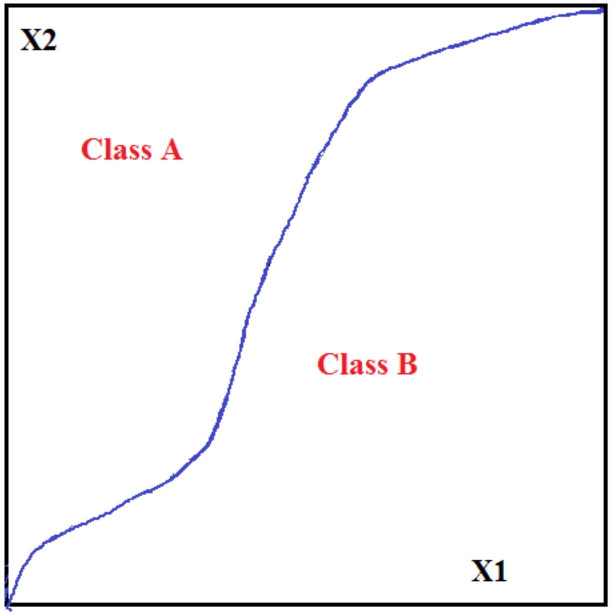
Distribution	Target	Link function	Offset	Weight
Poisson	Count	Log	Log(exposure)	1
	Rate(*)	Log	0	Exposure
Binomial	(#-yes, #-no)(^)	Logit	0	1
	Rate	Logit	0	Exposure
Gamma	Cost	Log	Log(exposure)	1
	Cost/exposure	Log	0	Exposure

*glm may give warning when target is not an integer

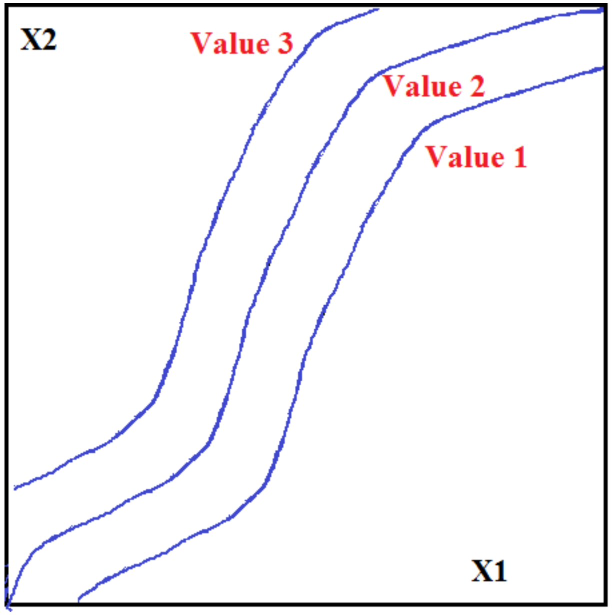
^ exposure is the same

What’s wrong with GLM

- ➢ GLM can only handle straight line (smooth curve) boundary (linear system)

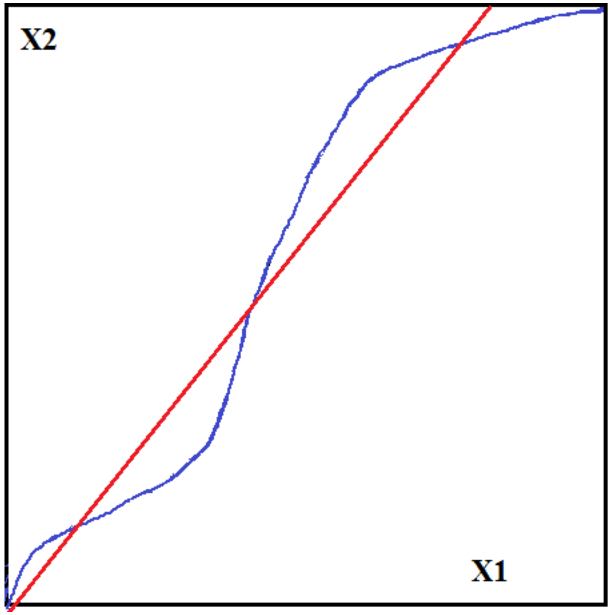


Classification



Regression

GLM approximation



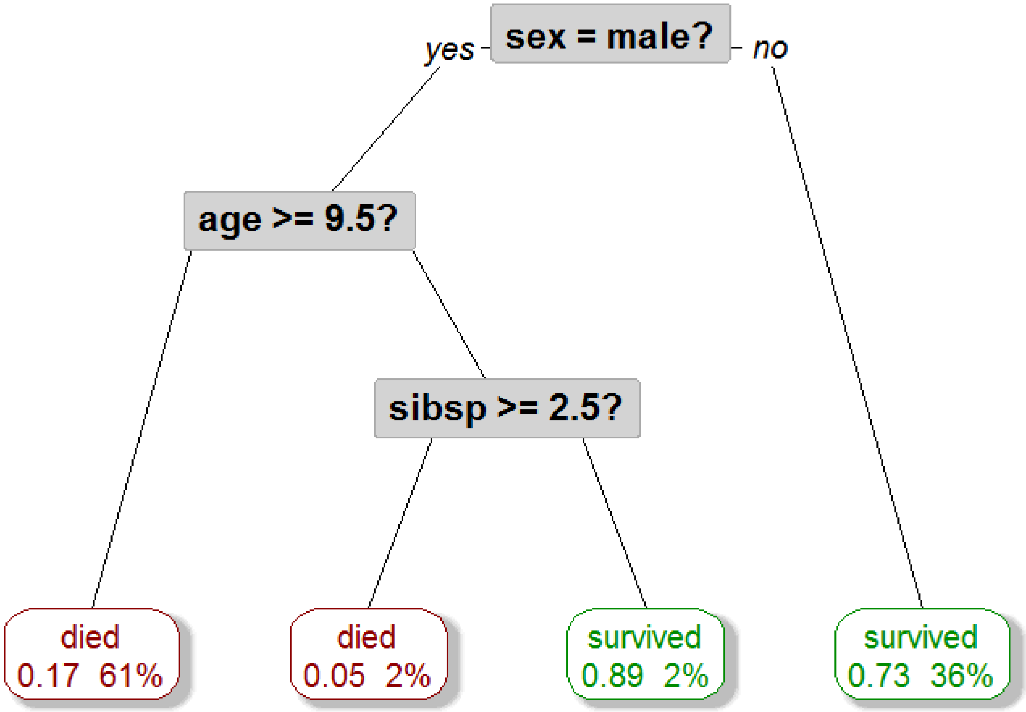
If there is a algorithm…

- Automatically analyze data & sift through all variables
- Separate relevant from irrelevant predictors
- No required variable transforms (logs, polynomial, etc.)
- Impervious to outliers & missing values
- Yield relatively simple & easy to understand models
- Required little to no intervention by analyst
- More accurate than logistic regression or other parametric tools

Too good to be true?

One example

- Titanic tragedy survivor analysis



- How could you do it?

Example - Characteristic CART Features

- Tree is relatively simple: focus is on limited variables
  - trees are often simple relative to problem
  - but they can be very large if necessary
- Accuracy is often high — about as good as any logistic regression on more variables
  - Not easy to develop a parametric model significantly better than a tree, unless it is linear structure
  - logistic regression will require intensive modeling effort
- Results can be easily explained to non-technical audience
  - More enthusiasm from decision makers for trees than for “black box”

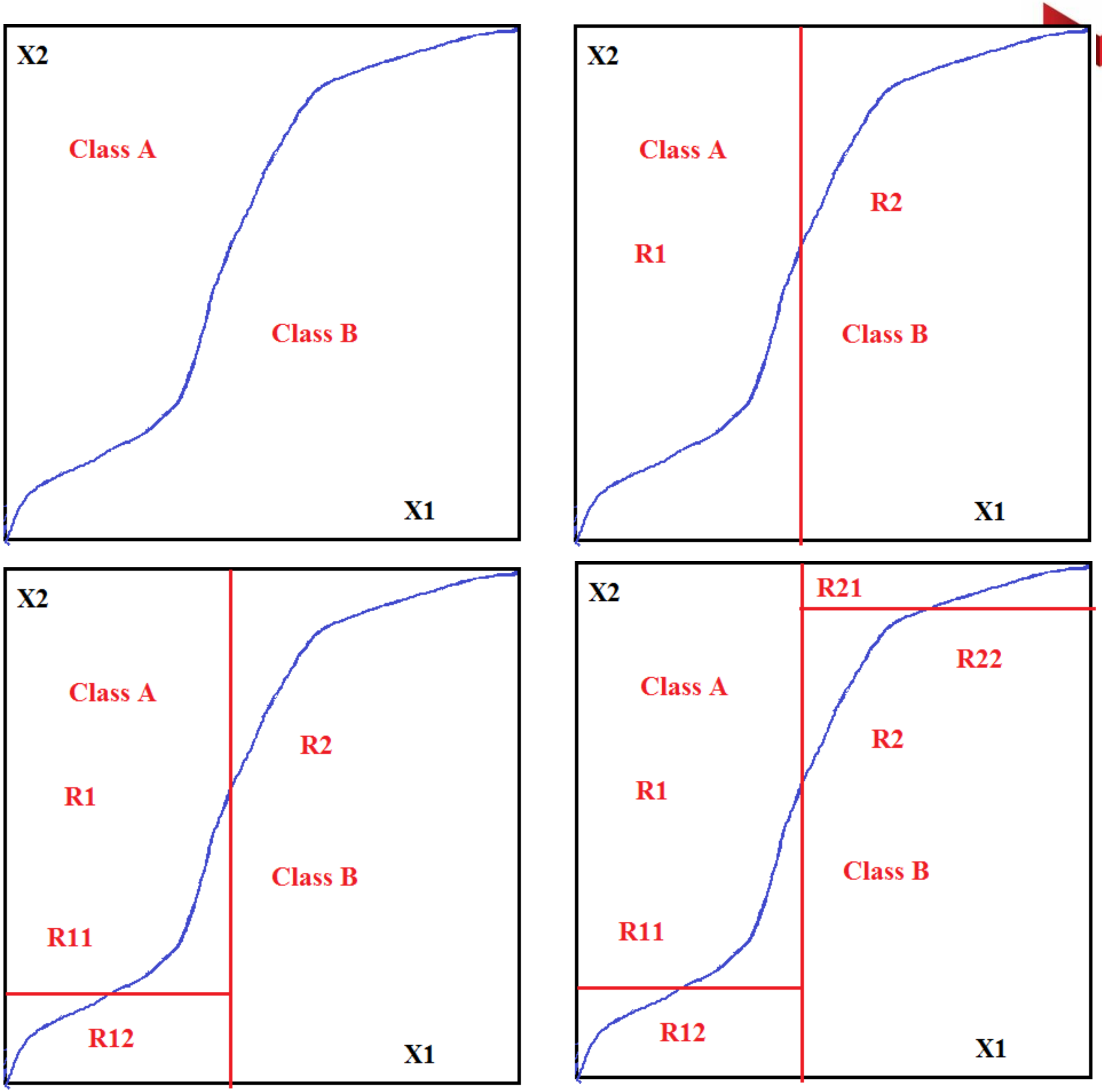
Decision Tree

- CART - Classification And Regression Trees
  - Commercial version
- Developed by Breiman, Friedman, Olshen, Stone in 70s/80s
  - Introduced tree-based modeling into the statistics
  - Rigorous approach involving cross-validation to select optimal tree
- There are many versions
  - CART - classic & popular
  - Software package variants (SAS, S-Plus, R…)
  - “rpart” package in R
- Data structure not apparent from a linear regression analysis

Decision Tree – Key Idea

Recursive Partitioning

- Take all data points
- Consider *all* possible values of *all* variables
- Select the variable/value **(X=t_1)** that produces the greatest "separation" in the target
  - **(X=t_1)** is called a "split".
- If $X< t_1$ then send the data to the "left"; otherwise, send data point to the "right"(generally speaking)
- Now repeat same process on these two "nodes"
  - Result is a "tree"
  - CART only uses *binary* splits
- Stop split data until certain criteria are meet


X2
Class A
Class B
X1
X2
Class A
R1
R2
Class B
X1
X2
Class A
R1
R2
Class B
R11
R12
X1
X2
R21
R22
Class A
R2
R1
Class B
R11
R12
X1

Two Core Questions

- How to find split points
  - Which variable among all
  - At which value or category
  - What criterion to use
- When to stop splitting
  - Avoid saturated model
  - Maximize accuracy, but not over-fitting - balance

Splitting Point

- Select the variable value (*X*=t_1) that produces the greatest "separation" in the target variable
- "Separation" defined in many ways
  - Different for regression & classification
  - Regression Trees (continuous target): use sum of squared errors

$$SSE_p = \sum_i (y_i - \mu)^2$$

$$SSE_c = \sum_i (y_i^L - \mu^L)^2 + \sum_i (y_i^R - \mu^R)^2$$

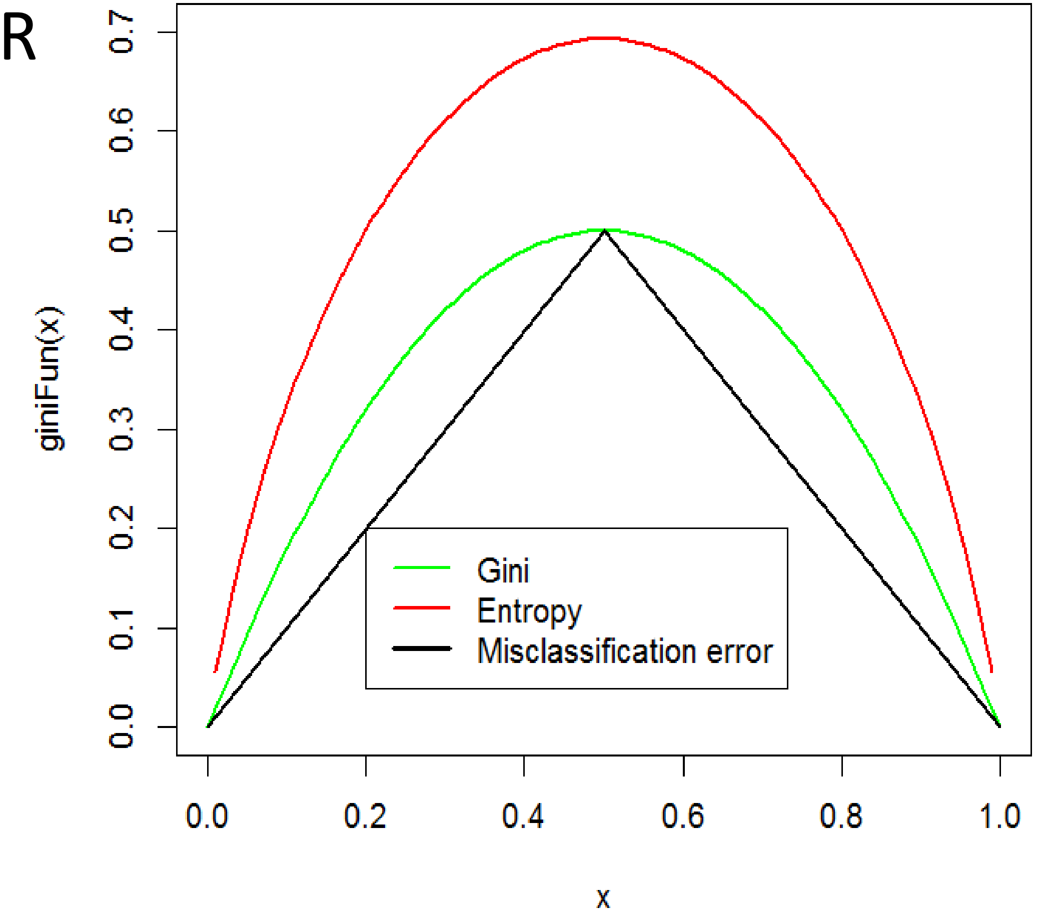
Select *X*=t_1 such that $\max\limits_{x_i,t}(SSE_P - SSE_C)$

  - Classification Trees (categorical target): choice of *entropy*, *Gini index*, "*twoing*" splitting rule.

Classification Trees

- Use measures of purity/impurity
- Intuition: an ideal tree model would produce nodes with only either class A or class B
  - Completely pure nodes
- Gini Index - purity of a node $f(p) = p(1-p)$
  - $f(p) = \sum_i p_i\,(1-p_i) = 1 - \sum_i p_i^2,\;\; p_i$ = freq of class i
- Entropy – information index $f(p) = -plog(p)$
  - $f(p) = \sum_i -p_i\,log(p_i) = -plog(p) - (1-p)log(1-p)$
- "twoing" - balance between purity & creating roughly equal-sized nodes
  - "twoing" is not available in the "rpart" package in R

- Max entropy/Gini when p=.5, no differentiation
- Min entropy/Gini when p=0 or 1, pure node
- Gini might produce small but pure nodes
- Plot in R



Surrogate Splits

Problem: if we have missing data on some predictor variables for an object, we don't know which class the object should be assigned to when that predictor is used for a particular split

Solution: we can use a similar split on another variable that is associated (correlated); we use these (surrogate) splits to assign the object to the class

Missing value can be solved in algorithm level

Titanic Example

By using R “rpart” package

Greedy Algorithm

- Define stopping criteria:
  - Stop when some minimum node size is reached
  - Or
  - Split only when decrease in cost, f(p), > a threshold
- Tree size
  - A very large tree may over fit the data
  - A small tree may not structure the important data structure

Cost Complexity Pruning

- Balance misclassification error vs. complexity
- Build a very large tree by using the greedy approach
- Define a sub tree by pruning nodes
  - Collapsing internal nodes by pruning leaves
- Find optimal T nodes at a given α that minimizes the cost-complexity criterion

$$C_{\alpha}(T) = R(T) + \alpha T$$

α called complexity parameter (CP), the complexity cost per terminal node

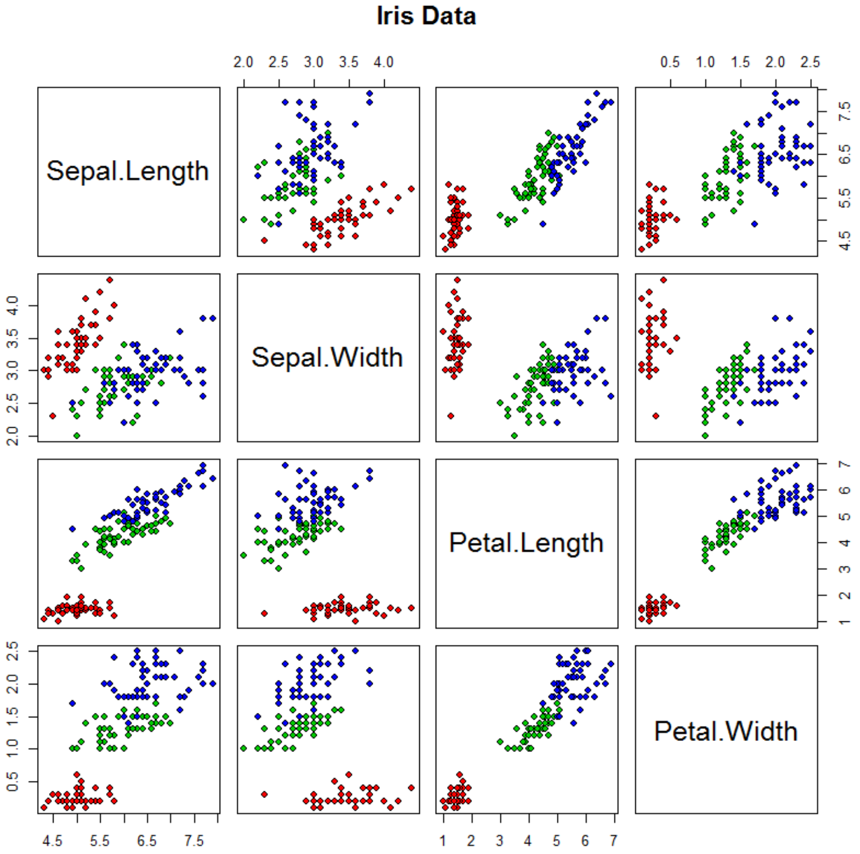
- When α is small, the penalty for having a larger tree is small, so T could be very large (extreme case → saturated model)
- When α increases, T decreases (extreme case → root only)

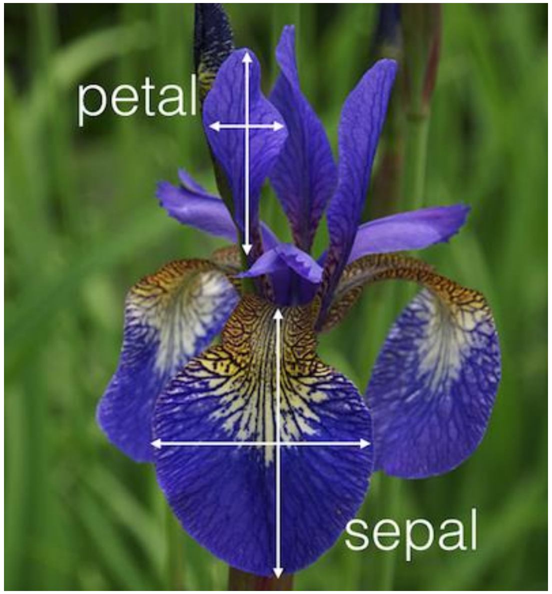
Find α by cross-validation

Example

summary(iris)

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```





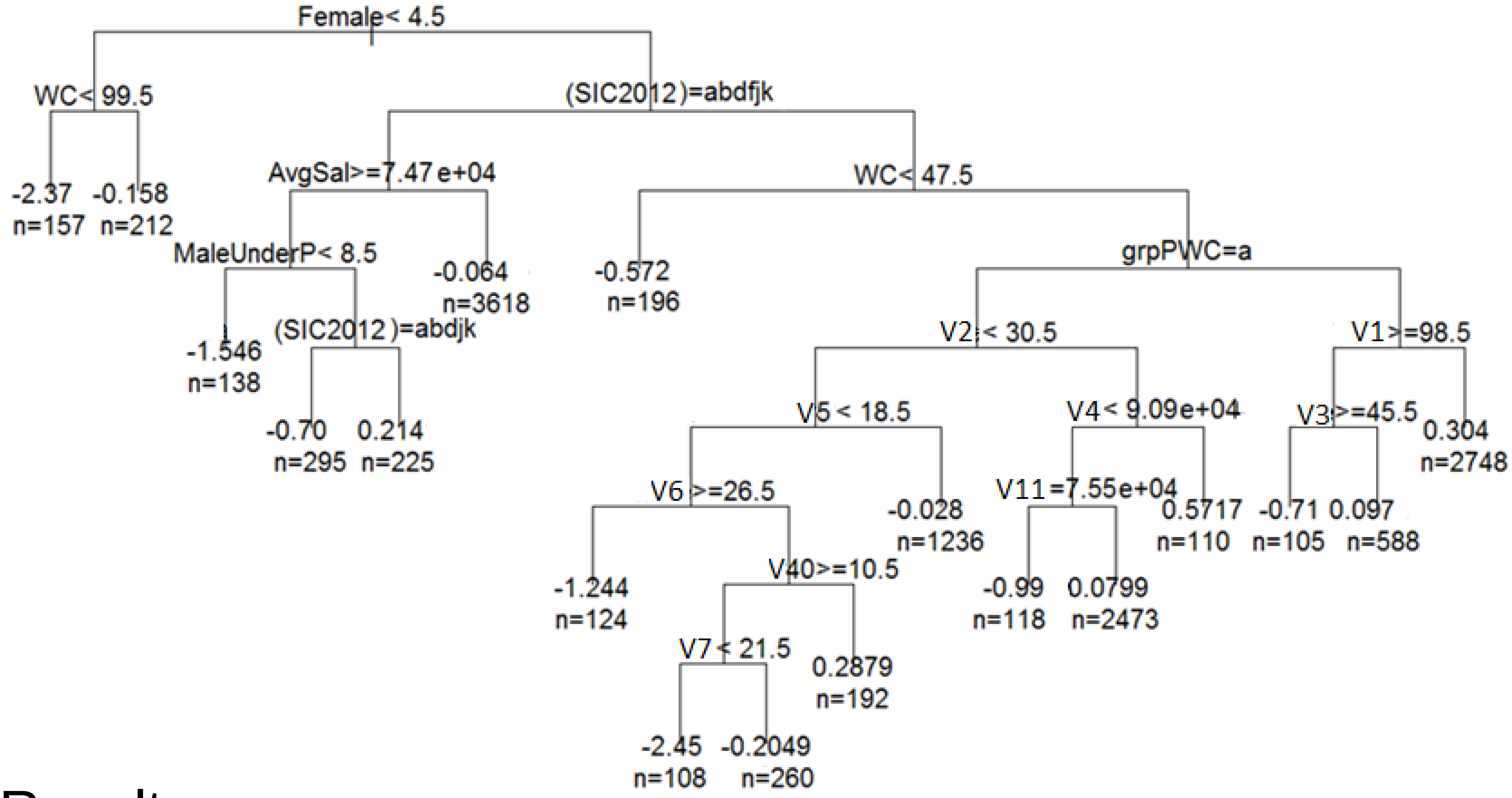
Classification vs. Regression Trees

- Splitting Criteria:
  - Gini, Entropy, Twoing
- Goodness of fit measure:
  - Misclassification rates
- Prior probabilities and misclassification costs
  - Available as model “tuning parameters”

- Splitting Criterion:
  - Sum of squared errors
- Goodness of fit:
  - Same measure
  - Sum of squared errors
- No priors or misclassification costs…
  - Just let it run

Insurance Application

- Business: US group Long-Term Disability(LTD)
  - ✓ About 13k policies, with lives per policies from 10 to 30k
  - ✓ Current pricing variables: about 30-40
  - ✓ Experience data of past 5 years with >80 variables
  - ✓ Major pricing variables: age, gender, industry, location, benefit structure
- Objective
  - ✓ To determine additional pricing variables and possible interaction terms (for pricing)
  - ✓ To identify groups with experience deviating from pricing assumptions (for UW)
- Client has experience with PM
  - ✓ Minimum efforts on business & data understanding
  - ✓ CART model vs. GLM model



Results

✓ Relatively easy to develop, interpret and understand

✓ Knowledge of business insights

✓ Implementation

Summary

Advantages

- No distribution assumptions on the variables
- Both classification & regression
- Excellent interpretability of tree structure
- Not significantly impacted by outliers in data
- Missing values handled at algorithm level

Disadvantages

- CART binary tree, may lead to instability
- Splits aligned with axes of feature space, may be suboptimal; very low efficiency in linear relationship

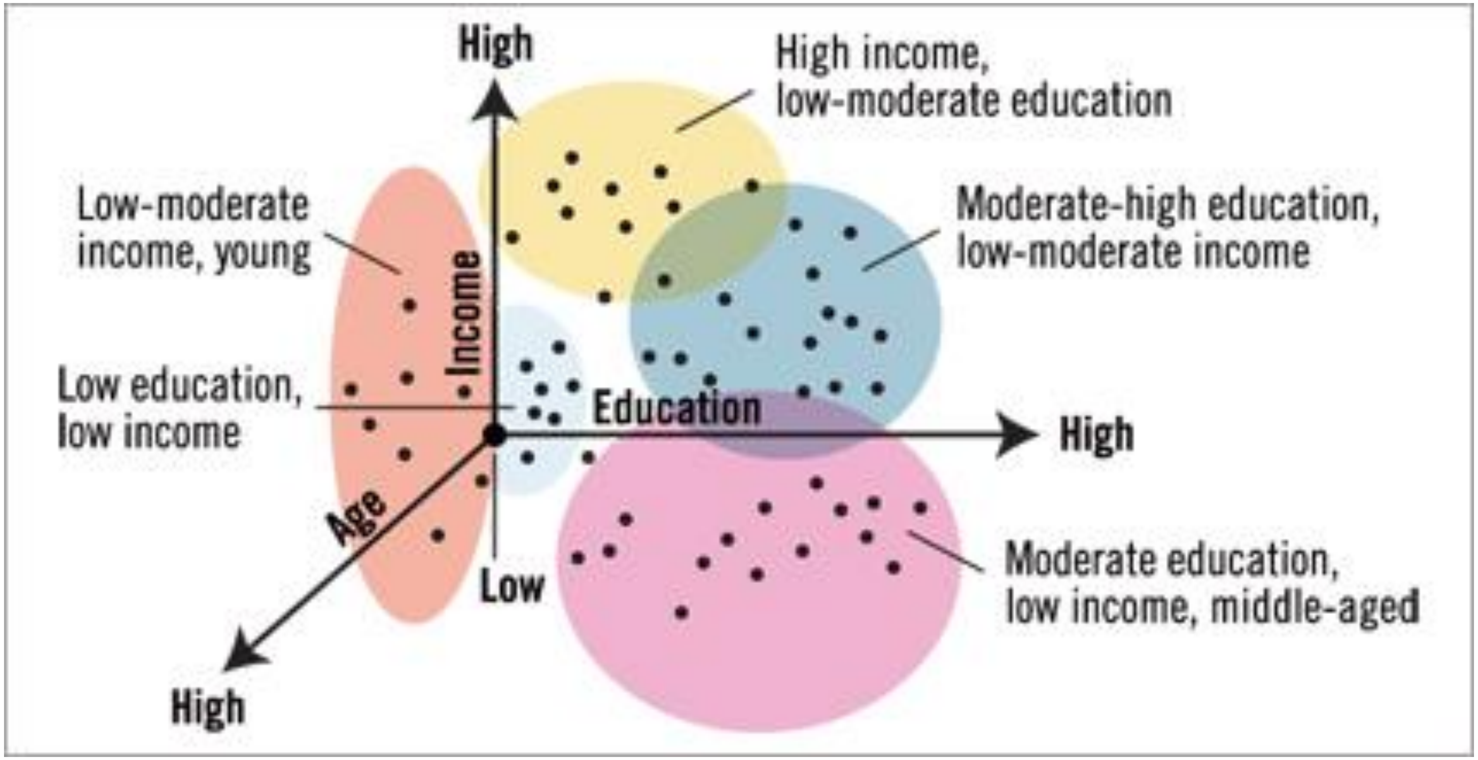
CART model

Break

Clustering

Clustering

- Given a dataset, group them into clusters such that
  - points within each cluster are similar to each other
  - points from different clusters are dissimilar
- Clustering is unsupervised classification: no predefined classes



Applications

- Biology – hierarchical classification in structure; genes sequence analysis
- Marketing - distinct groups in customer bases; to develop targeted marketing programs
- Climate- weather patterns in atmosphere & ocean
- Library - book categories in libraries
- Medicine - features & variations of disease; imaging
- WWW - group search results; social network
- Crime analysis: type, location
- City-planning: group houses by type, value, & location
- Earth-quake: cluster epicenters along continent faults
- Data compression: represent a whole cluster by index

Clustering Criterion

- Scalability ~ $n/n^2/n^3$
- Ability to deal with different types of attributes,
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise & outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability & usability

Algorithm

- Partitioning algorithms, K-measn/k-medoids
  - Maintain a set of k clusters; k known
  - Place points into their “nearest” cluster
- Hierarchical (Agglomerative)
  - Objects are more related to nearby objects than to objects farther away; objects are connected by distance
  - How to define “nearby” object
- Distribution-based: if objects belonging most likely to the same distribution; over-fitting issue
- Density-based: clusters are areas of higher density than the remainder of the data set
- Model-based

Partitioning Algorithms

- Construct a partition of a dataset of ***n*** objects into a set of ***k*** clusters
- Given a ***k***, find a partition of ***k*** clusters that optimizes the chosen partitioning criterion
  - *k-means*: Each cluster is represented by the center of cluster
  - *k-medoids* (PAM - Partition around medoids): Each cluster is represented by medoid which is the most centrally located object in a cluster

K-Means Algorithm

At a given *k*,

```
1. Select K points as initial centroids
2. Repeat
3.         Form K clusters by assign each points to its
           nearest centroid
4.         Re-compute the centroids of each cluster
5. Until centroids do not change
```

Distance Method

distance for two profiles X_i and X_j

Euclidean: $d(x_i, x_j) = (\sum_k (x_{ik} - x_{jk})^2)^{1/2}$ (p=2; L_2 norm)

easy to understand, but not scale invariant

Manhattan: $d(x_i, x_j) = \sum_k |x_{ik} - x_{jk}|$ (p=1; L_1 norm)

city-block distance

Chebychev: $d(x_i, x_j) = max_k |x_{ik} - x_{jk}|$ (p→∞; L_∞ norm)

Minkowski: $d(x_i, x_j) = (\sum_k (x_{ik} - x_{jk})^p)^{1/p}$

Pearson Correlation: $d(x_i, x_j) = 1 - r, r = cor(x_i, x_j)$

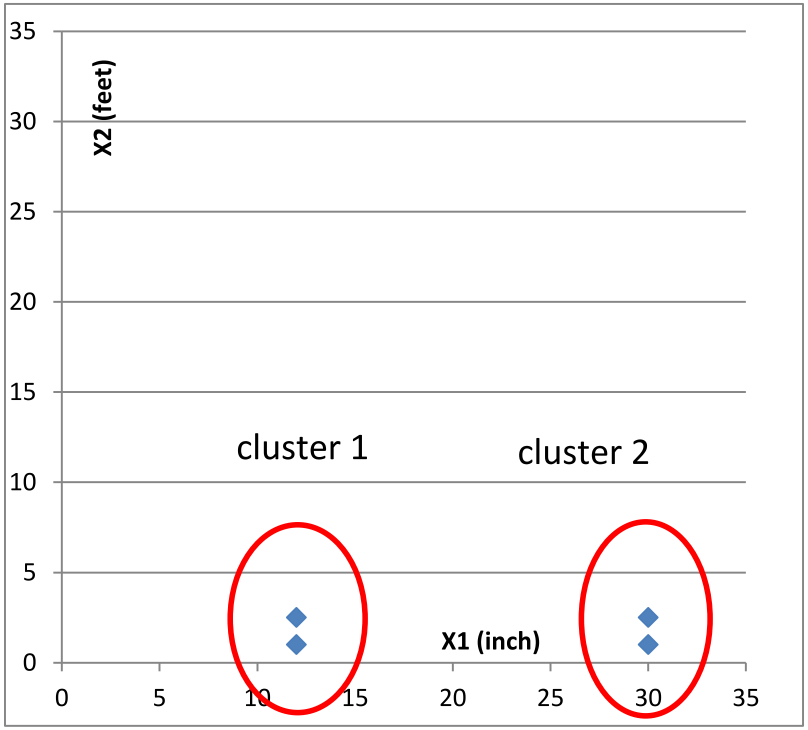
scale invariant but outlier sensitive

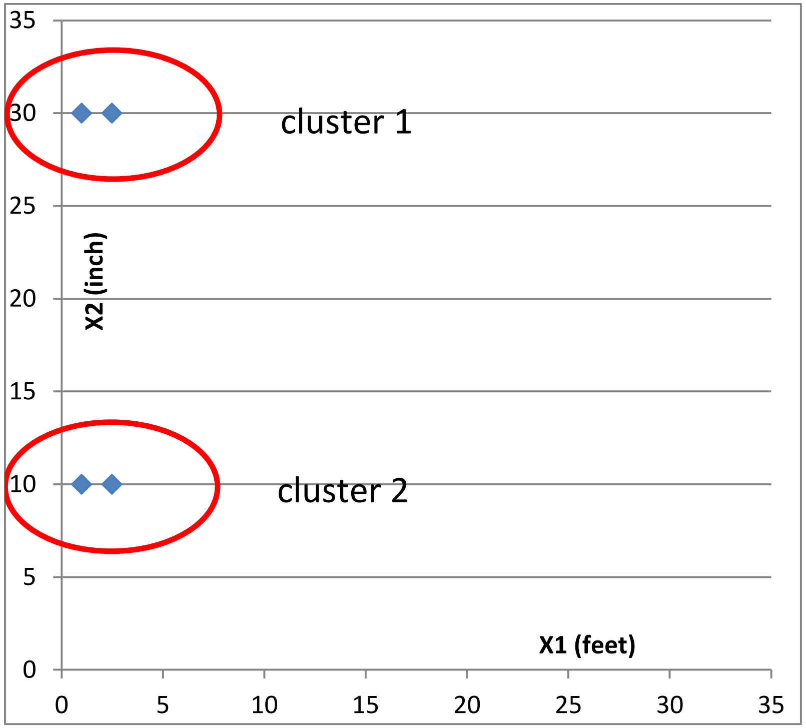
Spearman: Same as PCC but with ranked values

Others like Canberra, Jaccard, binary, ...

Standardization / Normalization

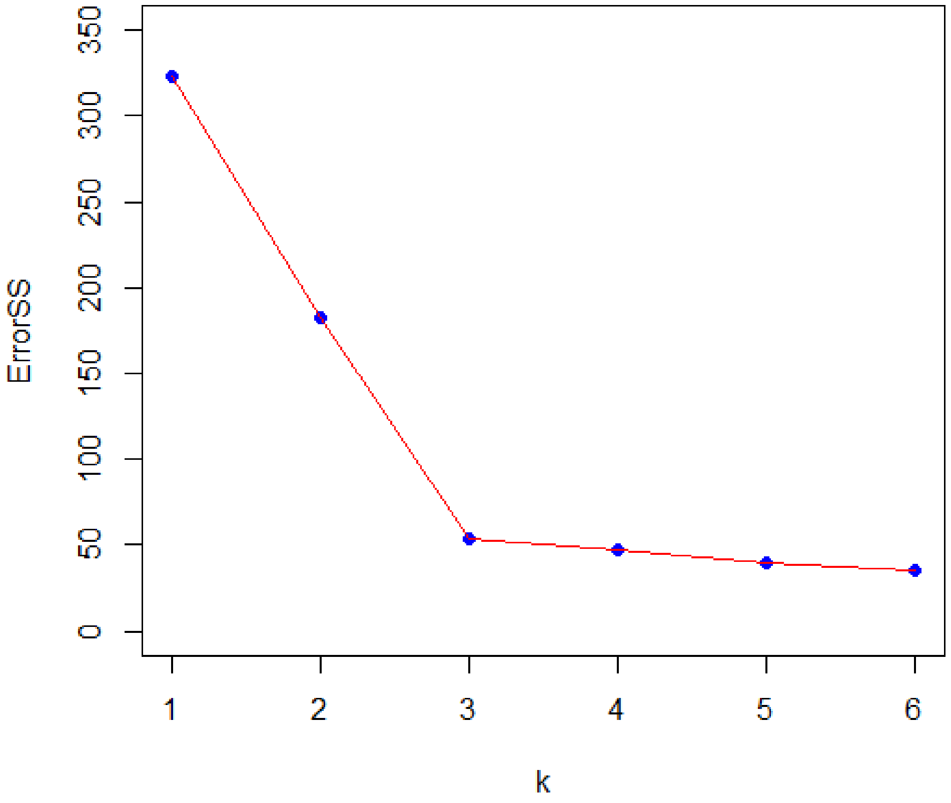
- Values of variables may have different units
  - Some will take vary large values, & "distance" can be large
  - Others may be small in values, & difference will be small
- Variable with high variability/range will dominate metric, & even lead to bias
- variable standardization / normalization



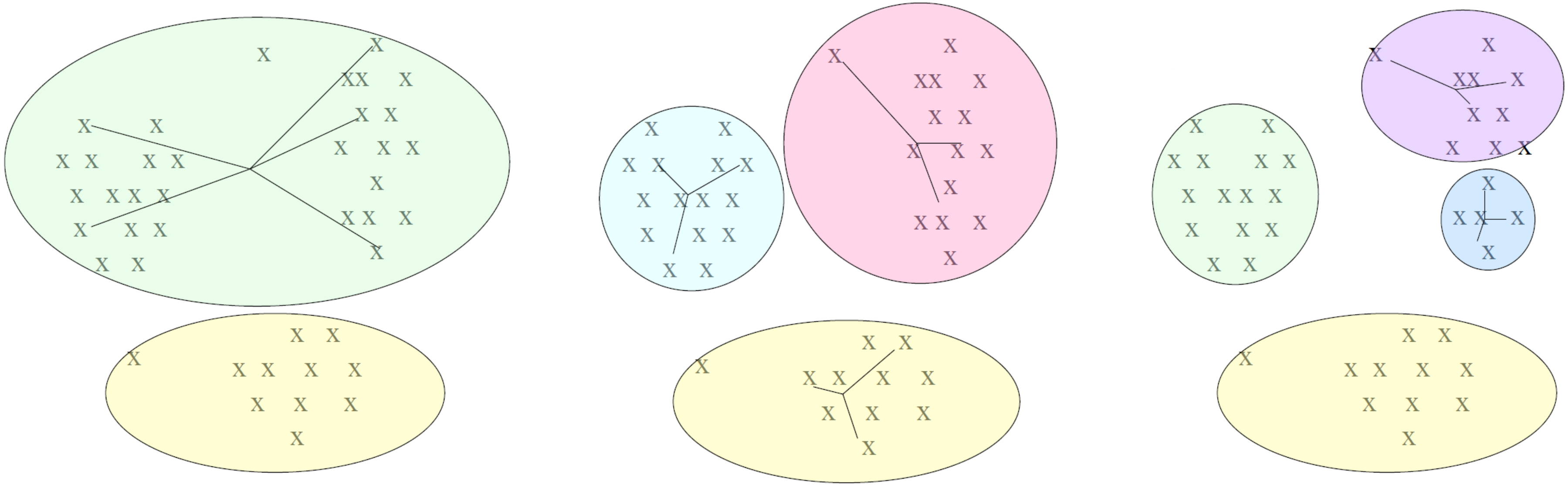


How to determine K

- If there are business reasons, go ahead
- Try different *k*, looking at the change in the average distance to centroid, as *k* increases
- Average falls rapidly until right *k*, then changes little



Interpretation of K

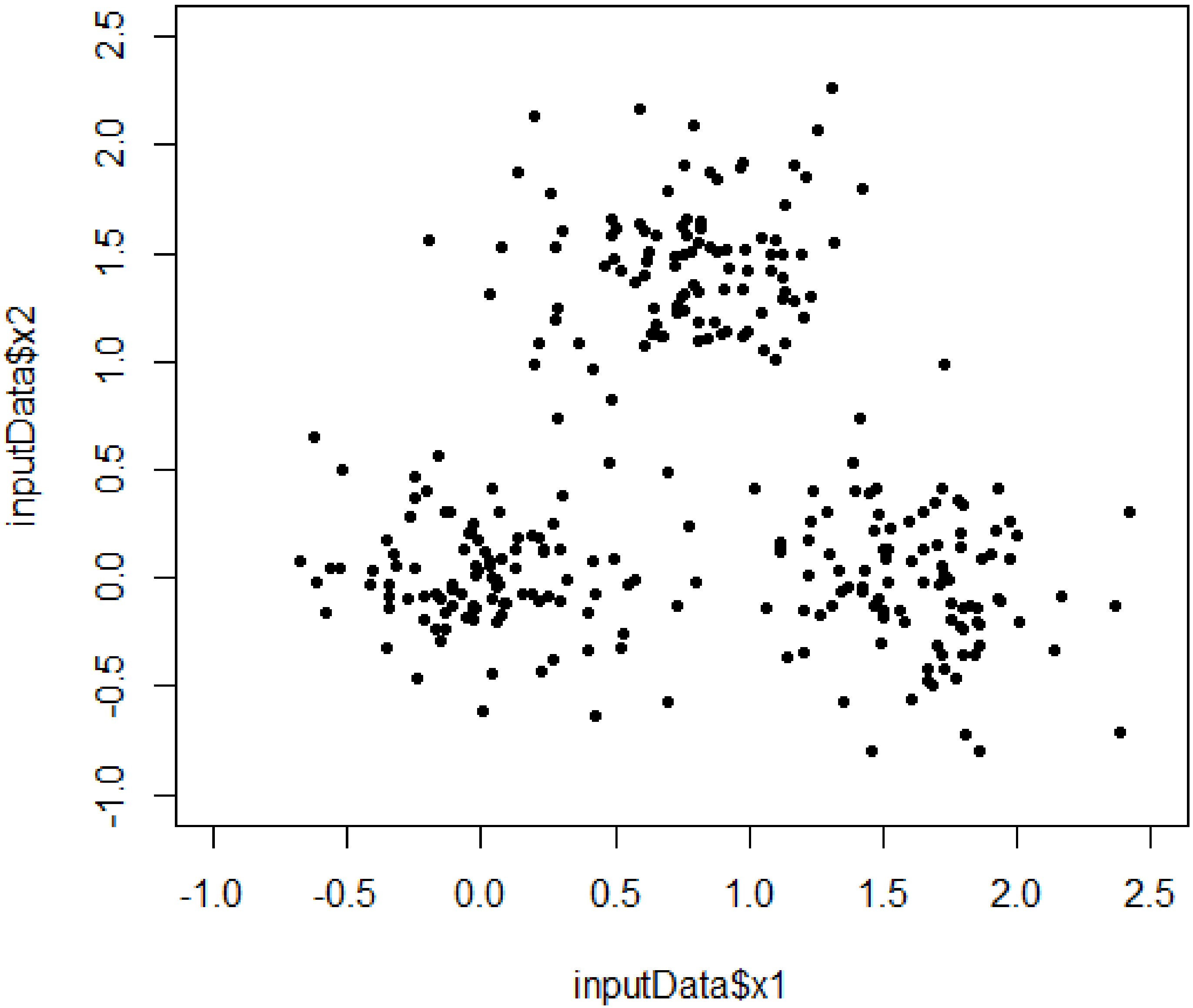


K=2
Too few; too much
Inter-cluster distances

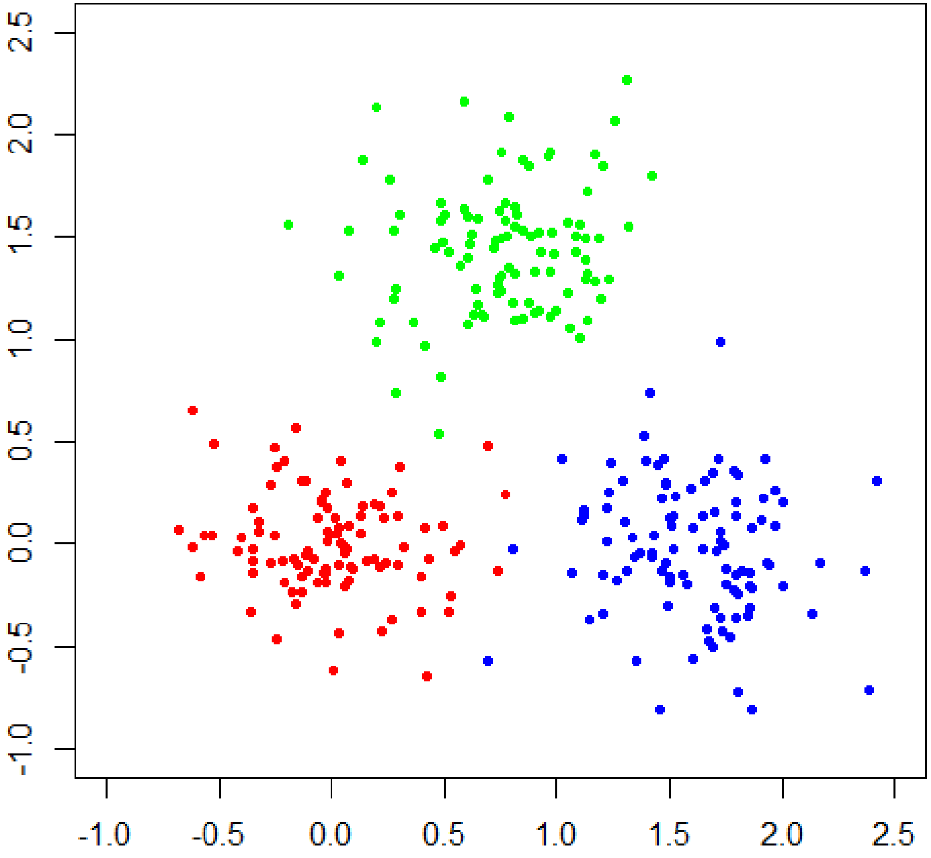
K=3
about right

K=4
Too many;
no significant improvement

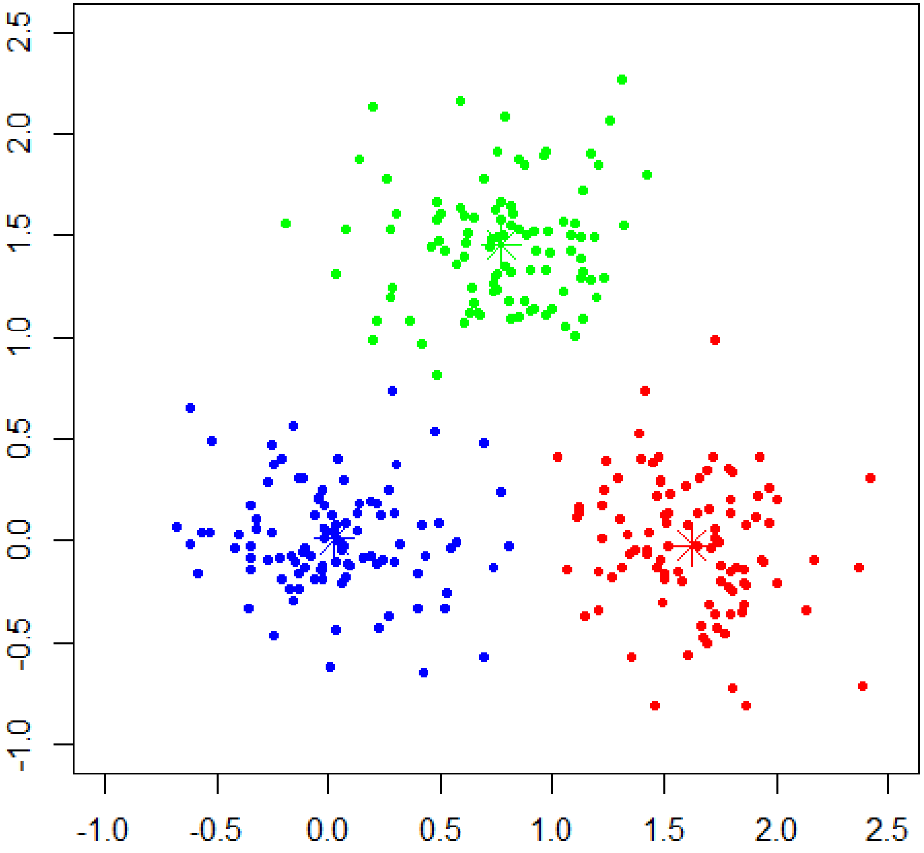
One example – demo in R



Original



Result



How could you do it in R? - example

Comments on *K-Means*

Strength: simple & *very efficient* ~$O(tkn)$, n = # objects, k = # clusters, t = # iterations, k, $t << n$.

Weakness

- Applicable only when *mean* is defined, what about categorical?
- Need to know *k,* the *number* of clusters, in advance
- Unable to handle noisy data & *outliers;* sensitive to outliers
- Not suitable for clusters with *non-convex shapes*
- *Sensitive to initialization*

A few variants of *k-means*

Selection of the initial *k* means; dissimilarity calculations; strategies to calculate cluster means

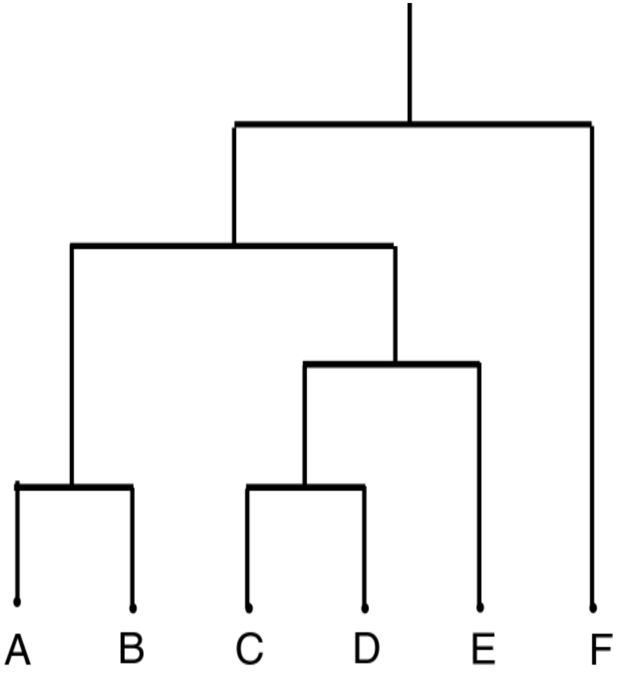
Hierarchical clustering

- Bottom up (aglomerative)
  - Start with single-instance clusters
  - At each step, join the two closest clusters
  - Design decision: define distance between clusters
- Top down (divisive approach / deglomerative)
  - Start with one universal cluster
  - Find two clusters
  - Proceed recursively on each subset
  - Can be very fast
- Both methods produce a *dendrogram*

Dendrogram

A tree data structure which illustrates hierarchical clustering techniques

- Each level shows clusters for that level.
  - – Leaf – individual clusters
  - – Root – one cluster
- A cluster at level i is the union of its children clusters at level i+1.



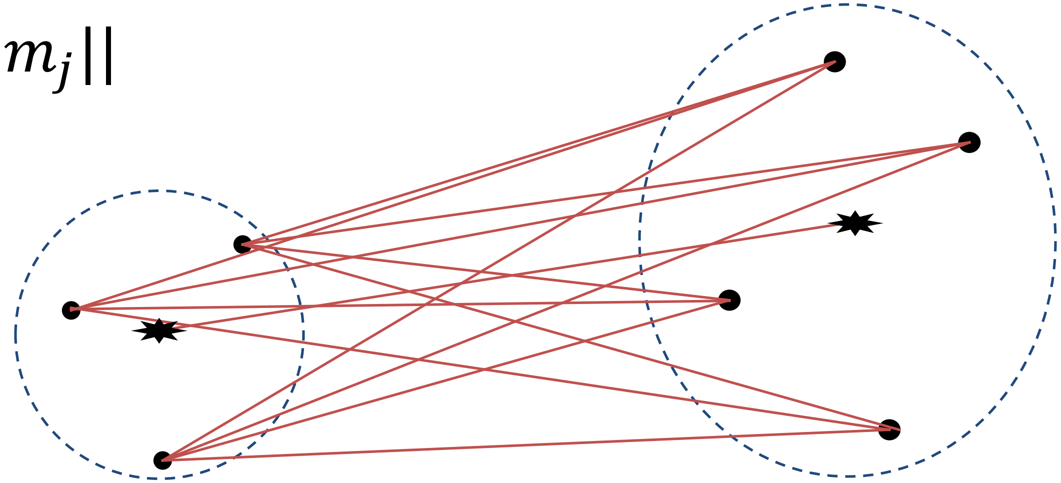
Hierarchical clustering - Agglomerative

Initially each data point in its own cluster
Iteratively clusters are merged together until one cluster

- Key Operation: repeatedly combine two nearest clusters
- Important questions
  - How do you represent a cluster of more than one point?
  - How do you determine the “nearness” of clusters?

Distance between clusters

- ***Single Link***: smallest distance between points
  $$d(C_i, C_j) = \min_{p_i \in C_i, p_j \in C_j} ||p_i - p_j||$$
- ***Complete Link:*** largest distance between points
  $$d(C_i, C_j) = \max_{p_i \in C_i, p_j \in C_j} ||p_i - p_j||$$
- ***Average Link:*** average distance between points
  $$d(C_i, C_j) = (1/n_i n_j) \sum_{p_i \in C_i} \sum_{p_j \in C_j} ||p_i - p_j||$$
- ***Centroid:*** distance between centroids
  $$d(C_i, C_j) = ||m_i - m_j||$$



An Example

Insurance Application of Clustering Analysis

By using
Hierarchical Clustering on Principle Components (HCPC)

Case Study: Risk Segmentation

- Foreign travel increased exponentially in past 50 years; associated risk impacted on life insurance

- There are mortality/morbidity differences between countries; location of residence can make large difference in mortality

- Objective
  - Assess foreign travel & residence risk
  - Compare all countries around the world on uniform basis
  - Data-driven conclusions based on facts and data, not popular opinion & preconceptions

Risk Segmentation

Life Expectancy (years)

Rank	Country	Life Expectancy	Rank	Country	Life Expectancy
1	Monaco	89.6	204	Chad	49.1
2	Japan	84.2	203	South Africa	49.5
3	Singapore	84.1	202	Guinea Bissau	49.5
4	San Marino	83.1	201	Swaziland	50.0
5	Andorra	82.6	200	Afghanistan	50.1
6	Switzerland	82.3	199	Central African Republic	50.9
7	Hong Kong	82.2	198	Somalia	51.2
8	Australia	82.0	197	Zimbabwe	51.5
9	Italy	82.0	196	Namibia	52.0
10	Liechtenstein	81.6	195	Gabon	52.2

Infant Mortality (deaths before age 1 per 1,000 live births)

Rank	Country	Infant Mortality	Rank	Country	Infant Mortality
1	Monaco	1.8	189	Malawi	77.0
2	Japan	2.2	190	Burkina Faso	78.3
3	Bermuda	2.5	191	Angola	81.8
4	Singapore	2.6	192	Niger	88.0
5	Sweden	2.7	193	Chad	91.9
6	Hong Kong	2.9	194	Guinea Bissau	92.7
7	Iceland	3.2	195	Central African Republic	95.0
8	Italy	3.3	196	Somalia	101.9
9	France	3.3	197	Mali	106.5
10	Spain	3.4	198	Afghanistan	119.4

Risk Segmentation

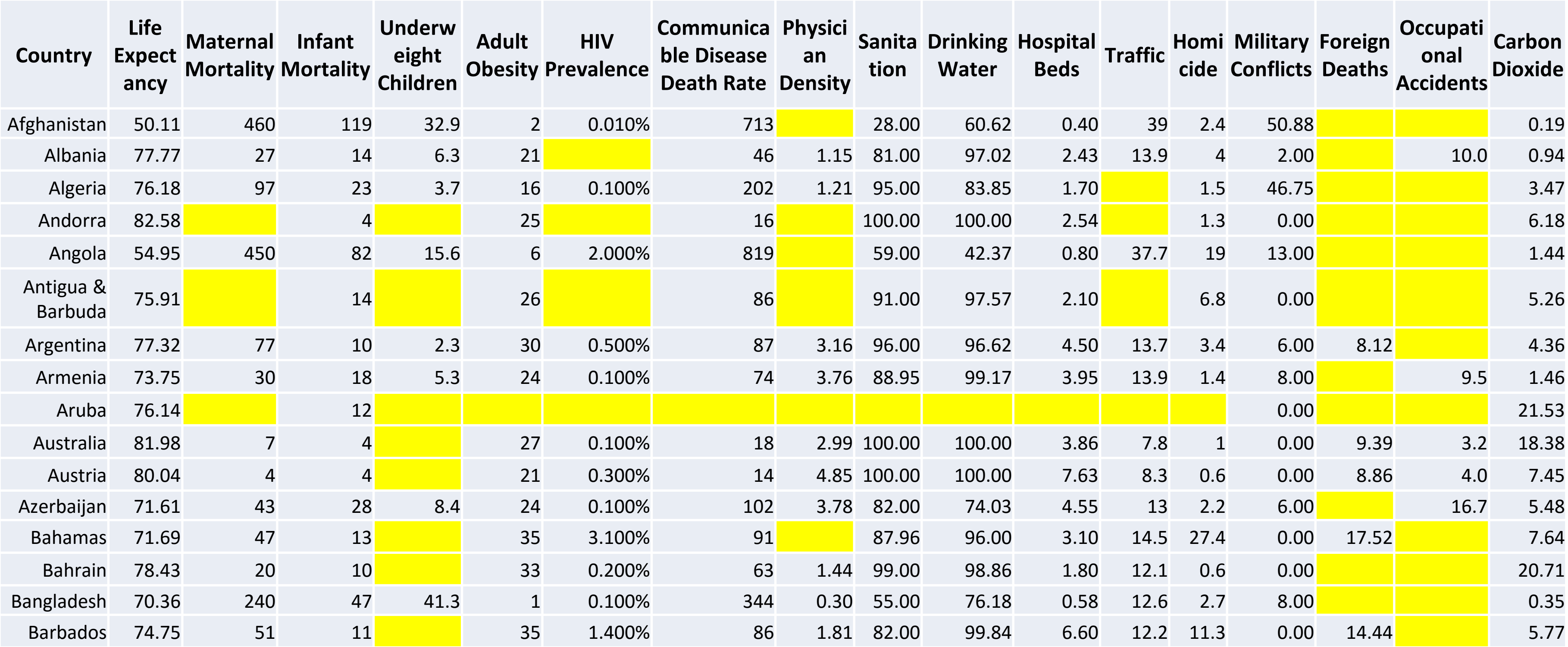
- Data for all 205 countries/regions
- 25 data fields to characterize foreign risks

  Life Expectancy(1), Maternal Mortality(2), Infant Mortality(3), Underweight Children(4), Adult Obesity(5), HIV Prevalence(6), Communicable Disease Death Rate(7), Physician Density(8), Sanitation(9), Drinking Water(10), Hospital Beds(11), Traffic(12), Homicide(13), Military Conflicts(14), Foreign Deaths(15), Occupational Accidents(16), Carbon Dioxide(17), Particulate Matter concentration(18), Internet Users(19), Mobile Phone(20), Road Density(21), GDP Per Capita (PPP)(22), Corruption(23), Education-Expected Years of School(24), Gini Index(25)

- Data sources

  CIA, WHO, World Economic Forum, World Bank, UN, Center for Systemic Peace, U.S. State Department, pueblo.gsa.gov, Elsevier, Transparency International

Risk Segmentation

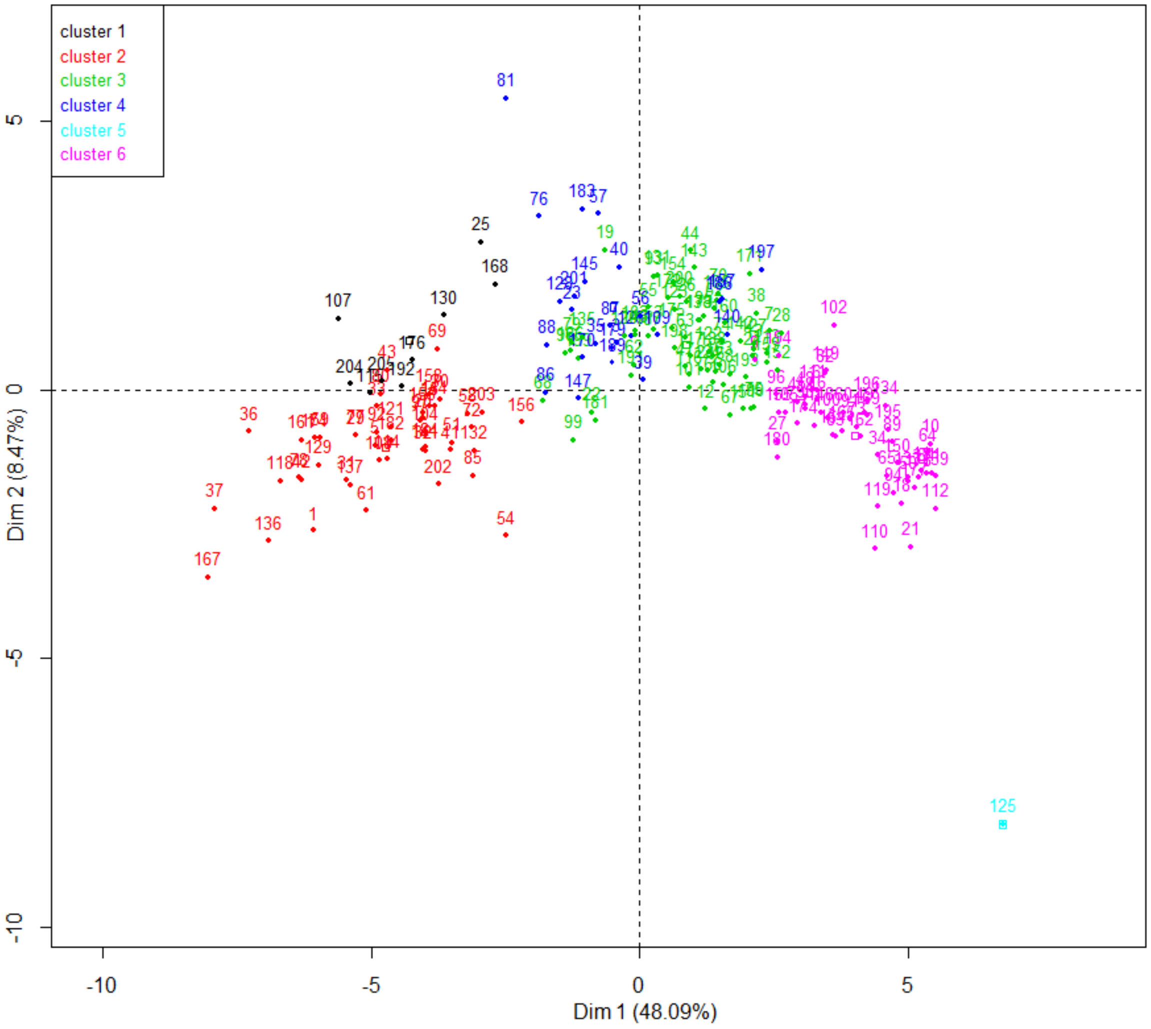


Country	Life Expectancy	Maternal Mortality	Infant Mortality	Underweight Children	Adult Obesity	HIV Prevalence	Communicable Disease Death Rate	Physician Density	Sanitation	Drinking Water	Hospital Beds	Traffic	Homicide	Military Conflicts	Foreign Deaths	Occupational Accidents	Carbon Dioxide
Afghanistan	50.11	460	119	32.9	2	0.010%	713	🟨	28.00	60.62	0.40	39	2.4	50.88	🟨	🟨	0.19
Albania	77.77	27	14	6.3	21	🟨	46	1.15	81.00	97.02	2.43	13.9	4	2.00	🟨	10.0	0.94
Algeria	76.18	97	23	3.7	16	0.100%	202	1.21	95.00	83.85	1.70	🟨	1.5	46.75	🟨	🟨	3.47
Andorra	82.58	🟨	4	🟨	25	🟨	16	🟨	100.00	100.00	2.54	🟨	1.3	0.00	🟨	🟨	6.18
Angola	54.95	450	82	15.6	6	2.000%	819	🟨	59.00	42.37	0.80	37.7	19	13.00	🟨	🟨	1.44
Antigua & Barbuda	75.91	🟨	14	🟨	26	🟨	86	🟨	91.00	97.57	2.10	🟨	6.8	0.00	🟨	🟨	5.26
Argentina	77.32	77	10	2.3	30	0.500%	87	3.16	96.00	96.62	4.50	13.7	3.4	6.00	8.12	🟨	4.36
Armenia	73.75	30	18	5.3	24	0.100%	74	3.76	88.95	99.17	3.95	13.9	1.4	8.00	🟨	9.5	1.46
Aruba	76.14	🟨	12	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	0.00	🟨	🟨	21.53
Australia	81.98	7	4	🟨	27	0.100%	18	2.99	100.00	100.00	3.86	7.8	1	0.00	9.39	3.2	18.38
Austria	80.04	4	4	🟨	21	0.300%	14	4.85	100.00	100.00	7.63	8.3	0.6	0.00	8.86	4.0	7.45
Azerbaijan	71.61	43	28	8.4	24	0.100%	102	3.78	82.00	74.03	4.55	13	2.2	6.00	🟨	16.7	5.48
Bahamas	71.69	47	13	🟨	35	3.100%	91	🟨	87.96	96.00	3.10	14.5	27.4	0.00	17.52	🟨	7.64
Bahrain	78.43	20	10	🟨	33	0.200%	63	1.44	99.00	98.86	1.80	12.1	0.6	0.00	🟨	🟨	20.71
Bangladesh	70.36	240	47	41.3	1	0.100%	344	0.30	55.00	76.18	0.58	12.6	2.7	8.00	🟨	🟨	0.35
Barbados	74.75	51	11	🟨	35	1.400%	86	1.81	82.00	99.84	6.60	12.2	11.3	0.00	14.44	🟨	5.77

- Main challenges
  - Many missing values
  - Different weights on certain fields
    - ✓ For example, life expectancy

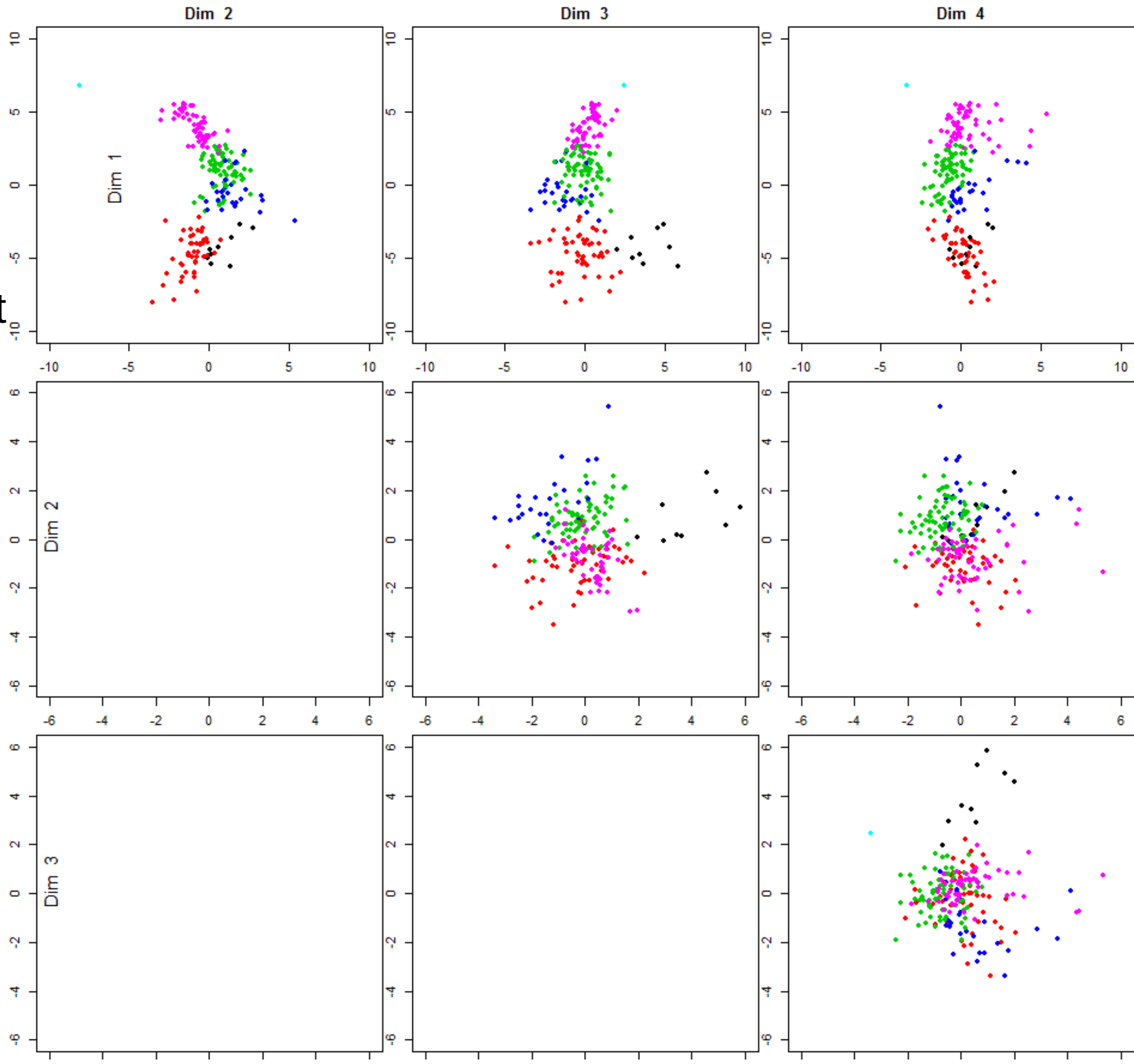
Risk Segmentation

- Missing values are dealt with at algorithm level
- PCA analysis followed by hierarchical clustering
  - Principle Component Analysis – explain variance in data
  - Weights are based on judgment, and considered at hierarchical clustering
- Results on 6 clusters
  - Number of clusters is a free parameters

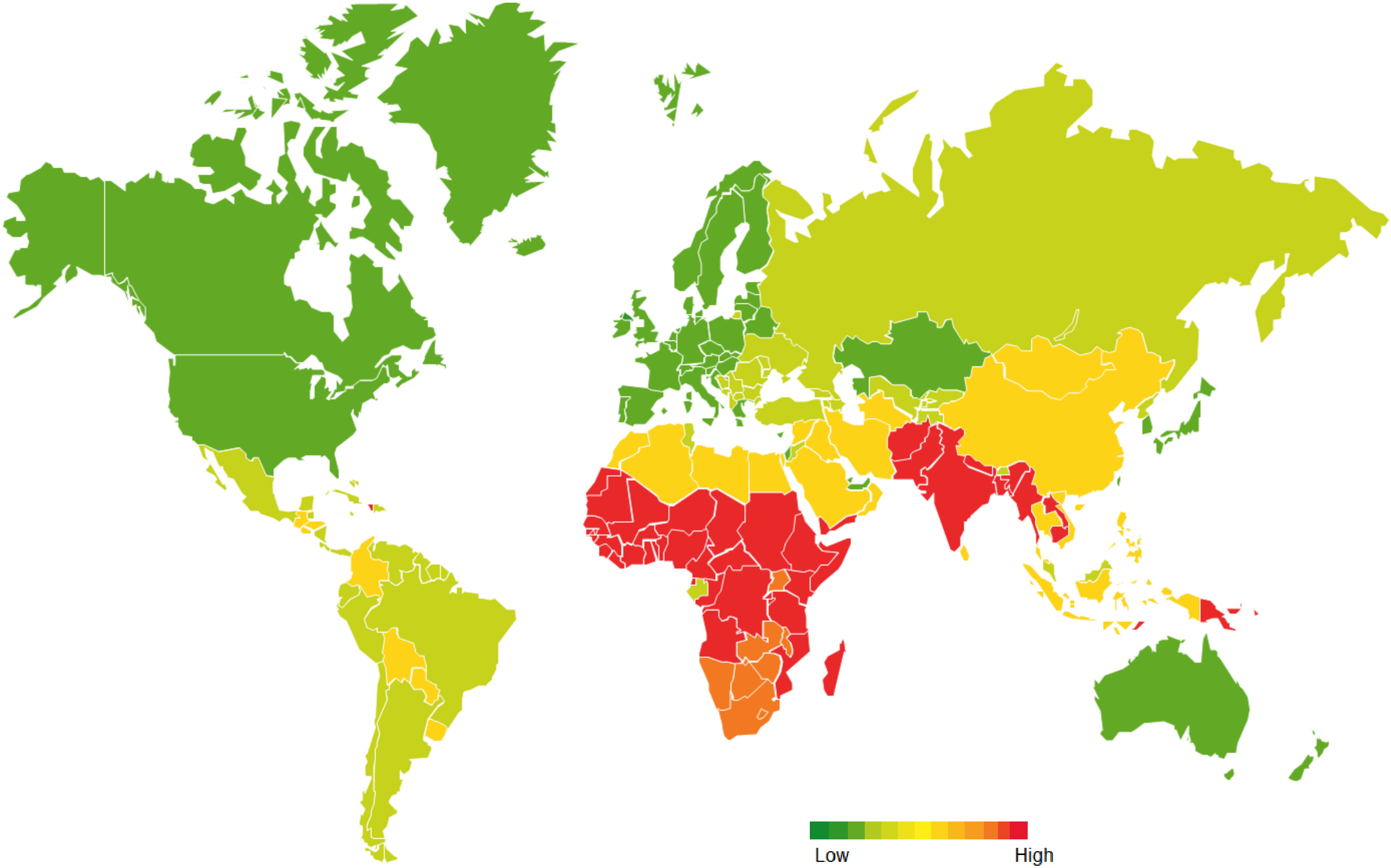


Risk Segmentation

- Data visualization
  - Scatter plot of first 4 components
- Component 1
  - ✓ Life expectancy, Infant mortality, Sanitation, Education Expected Years Of School, Internet users
- Component 2
  - ✓ Underweight children, Gini index, Foreign death, road density



Risk Segmentation



Comments on Hierarchical Clustering

Strengthens

- Include categorical variables
- May be required; Hierarchy results easy to understand
- Better clustering quality

Weakness

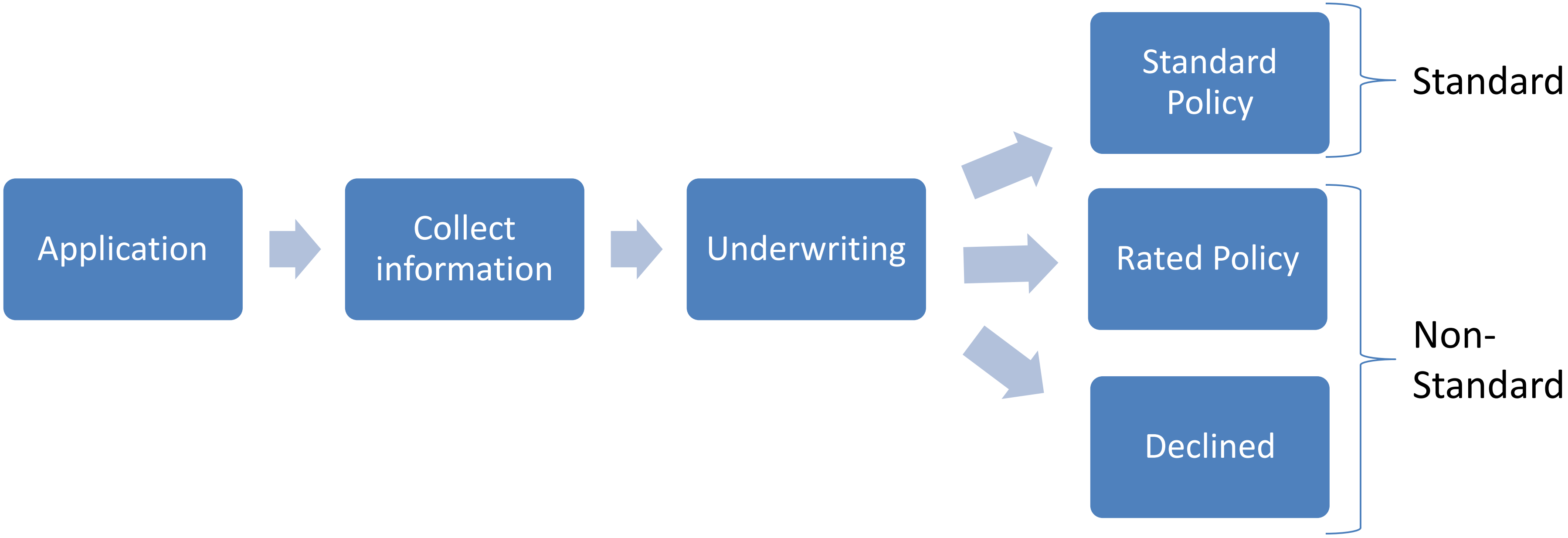
- Not saleable well: time complexity >$O(n^2)$,
- Not be able to undo what was done previously

MARYVILLE UNIVERSITY

Final Project

Goal: Using PM to solve business problem

- Ideally, predict individual mortality & offer product
  - But to do that, what do we need?
- Alternatively, predict UW decisions
  - What are advantage and disadvantage?



Final deliverables

- Complete model with proofed performance
- A comprehensive document to client; your target audience will be mid-/high-level mangers in bank/life, & actuaries in life product development/pricing/valuation
- A presentation is also required (to both groups)

Logistics

- A team of two students
  - Group 1 = Eric / Hui; Group 2 = Haifeng / Anlin; Group 3 = Nick / Mengchu; Group 4 = Jie / Luke; Group 5 = Alina / Matthew"
- Both are required to make equal contributions to final deliverables; each have 10-15 min presentation, focusing on different aspects of the project
- Project paper is due right before the last class

Grade

- Total 30 points toward the final grade
- Modeling 10; Results 4; Report 8; Presentation 8

Considerations

- Why we have to do this PM project
- How we come to the final model; what the major issues we have to deal with when we build the model
- How we select variables with predictive power; any issue
- What we learn from the model
- How we would implement it; what are required from client
- How we would handle anti-selection; any method to mitigate
- What the key indicator that PM project will be successful in real business

On last week class

- A full report for both managers & actuaries
  - Introduction; Background/Objective; Data/Data Process; Modeling/Procedure; Performance/Results; Implementation; Appendix/Attachment (if necessary)
- Model, including R script
  - Repeatable results for actuary to replicate
  - If other tools (e.g. Excel) used, please include
- Presentation
  - Separate presentations, 10-15 min each

Data

- Missing values
- Grouping
- Any question? Do not wait until last minute